METADATA 101

"data about data"

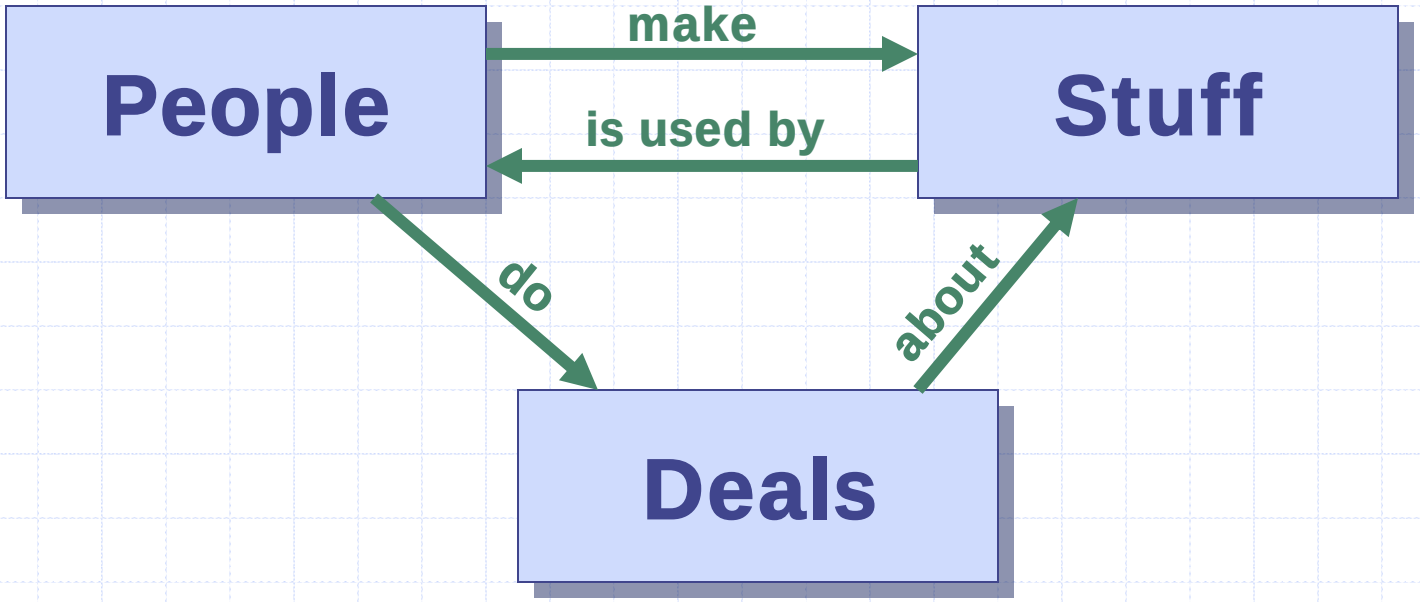
Sam LeFevre (slefevre@utah.gov)
Environmental Epidemiology Program
June 25, 2008

Objectives

- This presentation is an overview of descriptive metadata, using the Utah Environmental Public Health Tracking Network metadata as examples.

- This presentation
  - Reviews the history of metadata
  - Shows examples of common metadata used in everyday life
  - Defines the different categories of metadata
  - Emphasizes standards
  - And provides some discussion about Utah's implementation of descriptive metadata based on lessons learned by the tracking network so far.

Conceptualizing!



Some historical developments

- 1789: Card catalog first conceptualized in Paris
- 1960: Machine Readable Cataloging (MARC) standards implemented
- 1980: CERN develops Enquire, a first tag-based data protocol (lead to HTML in 1991 and XML in 1998)
- 1989: World wide web started, data sharing moved forward

Some historical developments

- 1990: Federal Geographic Data Committee (FGDC) organized
- 1998: Dublin Core, the first standard for digital data (http://dublincore.org/)
- 2002: Metadata object description schema implemented (http://www.loc.gov/standards/mods/)
- 2008: ISO standards for metadata completed

Kinds of metadata

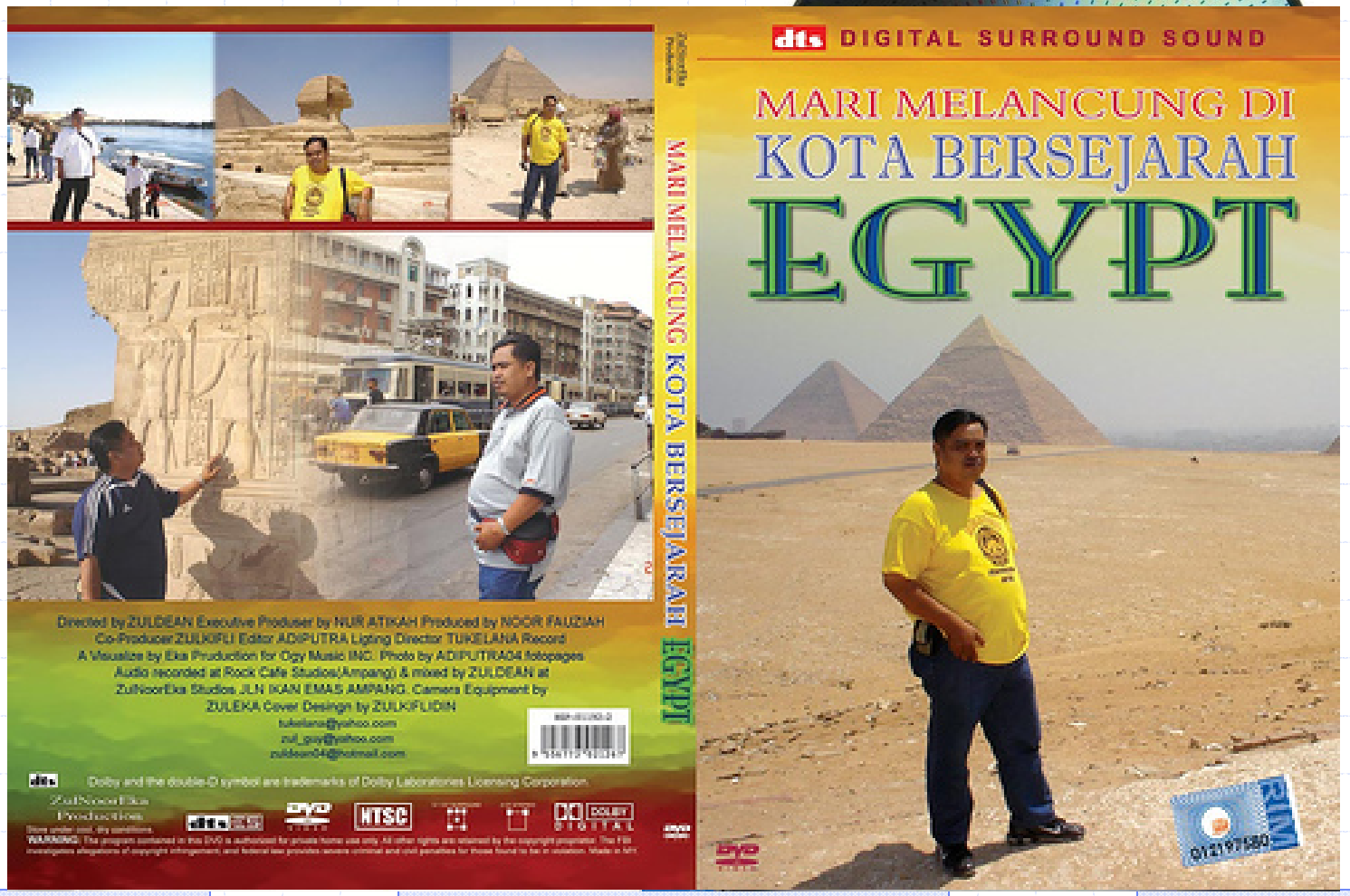
- Descriptive (Administrative)
  - Classification, cataloging, discover, …
- Administrative
  - Retrieval, re-use, version tracking, …
- Technical
  - Processing instructions (e.g., data imbedded in a video streams)
- Structural
  - Table schema, data dictionary, filed-value thesauri and vocabulary, …
- Messaging
  - Tag words, format instructions, …


```
- <File xmlns="http://TestMap2.FlatFileSchema1">
  - <OuterLoop xmlns="">
    - <ARecord>
        <AData>adata</AData>
      </ARecord>
    - <ZRecord>
        <ZData>zdata</ZData>
      </ZRecord>
  </OuterLoop>
```

Example

DATA

Metadata



Example

HO_UBLR_MD_2007.xml
XML Document
22 KB

HO_UBLR_STD_2007.sas
SAS Program
1,816 KB

HO_UBLR_STD_2007.xls
Microsoft Excel Worksheet
20 KB

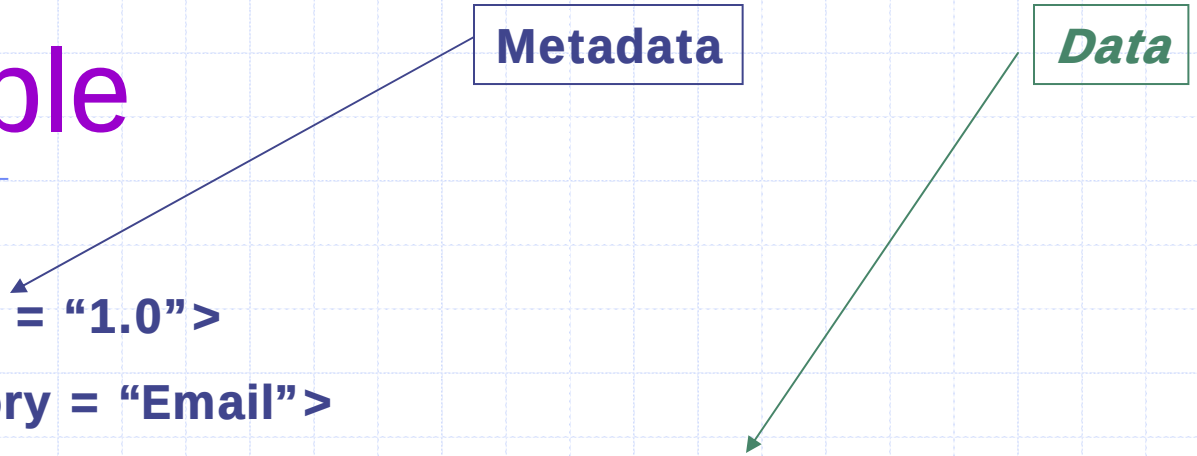
HO_UBLR_MD_2007.doc
Microsoft Word Document
58 KB

Example

DATA

Metadata

Example

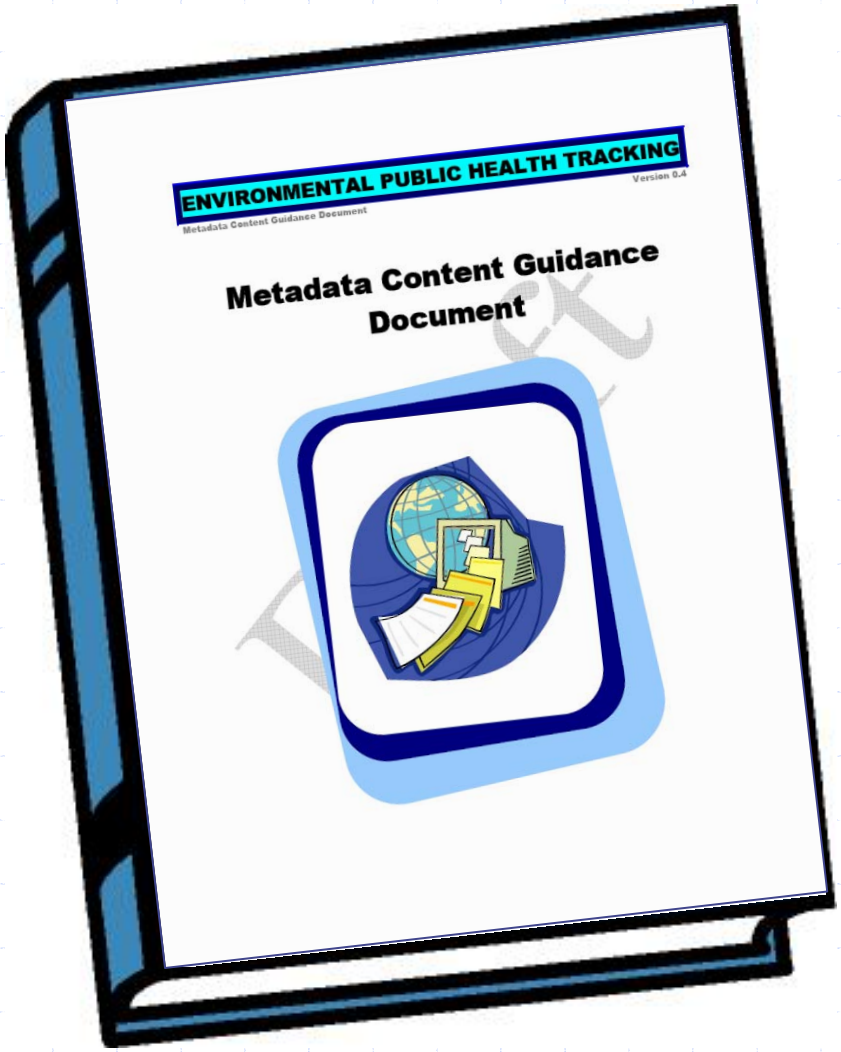



```
<xml version = “1.0”>
<note category = “Email”>
 <date format = “mmddyyyy”> 06/25/2008 </date>
 <to> Sam </to>
 <from> Wu </from>
 <heading> Reminder </heading>
<body> Don’t forget to come to the brownbag today </body>
</note>
</xml>
```

Example

A Data Element	**84116**
ZIP Code	The data element's name
Definition	The unique identifier of a postal district or post office
Data Type	Character
Length	5
Allowable Characters	0, 1, 2, 3, 4, 5, 6, 7, 8, 9

Example



Metadata on metadata!

Uses (Business Cases)

- A means to catalog, organize and track data (or anything else that can be described)
- A means to describe and track the provenance, version, completeness and processing of data
- A means of advertising data and informing potential users about the data

Best practices

- Old message:
  - **If you create metadata other people can discover your data.**
- New message:
  - **If you create metadata you can find your own data.**

Uses (Business Cases)

- Provides a citable document which allows data owners to be credited for the uses of their data
- A means to share data regardless of platform
- A tool to support authenticity of digital data
- A tool for data accountability
  - Repeatable processes
  - Defensible processes

Content of Metadata

- Defined by Standards
  - Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata (www.fgdc.gov)
  - Dublin Core Metadata Initiative (www.dublincore.org)
  - ISO 11179: Metadata Registry
  - ISO 19115: Geographic Metadata
  - ISO 19139: Geographic Metadata XML

  (www.iso.org)

  (metadata-standards.org)

Possible disadvantages

- Metadata is still not well defined
  - EPHTN white paper
  - PHIN VADS
- May be mis-used
  - Potentially reveals sensitive practices or processes
  - Cyber-crime, cyber-terrorism
  - Public awareness
- Requires time and resources to create, synchronize and maintain
- Increasingly difficult to formulate with increasing relational complexity of the data

Descriptive Metadata

- Provides Descriptive Information
  - Title, Abstract, Purpose
- Provides Administrative Information
  - Owner, Provenance, Version, Coverage, Completeness
- Provides Use Information
  - Security, Access, Constraints, Liability
- Provides Some Structure Information
  - Entity and attribute descriptions
  - Data dictionary

Example

- **Title: Utah Blood Lead Registry - 2007**
- **Abstract: Blood lead test results on all Utah residents reported to the Environmental Epidemiology Program from January 1, 1996 through December 31, 2007.**
- **Purpose: The Utah Blood Lead Registry tracks Utah residents with elevated blood lead levels to ensure they receive appropriate medical treatment and provides surveillance data for epidemiologic investigation of blood lead poisoning in Utah.**
- **Supplemental Information: Elevated blood lead test results are reportable under Utah Rule R386-703, Injury Reporting Rule. Other test results are reported voluntarily. The UBLR includes data on children and adults. Medicaid required testing of enrolled children. Other children tested based on history of exposure. Adults are tested on history of occupational exposure. Data are reported in both electronic and paper formats. Source data are maintained by the EEP for five years.**

Description

Should include:

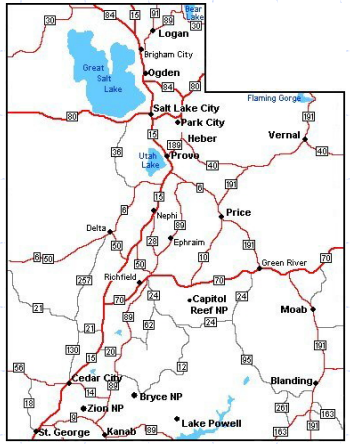
- **The descriptive name of the dataset**
  - **Descriptive: Utah Blood Lead Registry – 2007**
  - **Logical: HO_UBLR_STD_2007.sas**
- **Designation of the subject of the data**
- **General description of the coverage (demographic, geographic, temporal, diagnostic, etc.) (and scales) of the data**
- **Original purposes or uses of the data**

Coverage

- Administrative data elements used to detail coverage
  - Free Text
    - Spatial Domain: The Great State of Utah (the Beehive State)
  - Internal (Metadata Reserved Word) Standard Reference
    - Spatial Domain
      - Scale: Statewide
  - External Standard Reference
    - Spatial Domain
      - Thesauri: Geographic Names Information Standard (GNIS)
      - Key Word: Utah
      - Key Word: UT
  - Metadata Attribute
    - Spatial Domain
      - West Bounding Coordinate: -144.042925
      - East Bounding Coordinate: -109.041501
      - North Bounding Coordinate: 42.001718
      - South Bounding Coordinate: 36.997693

Content of Metadata

- Thesauri
  - Geographic Name Information System (GNIS)
  - Federal Information Processing Standards (FIPS)
  - US Postal Service Publication 28
- Vocabularies
  - Utah
  - UT
  - 49
- Free Text
  - Beehive state
  - Life elevated state



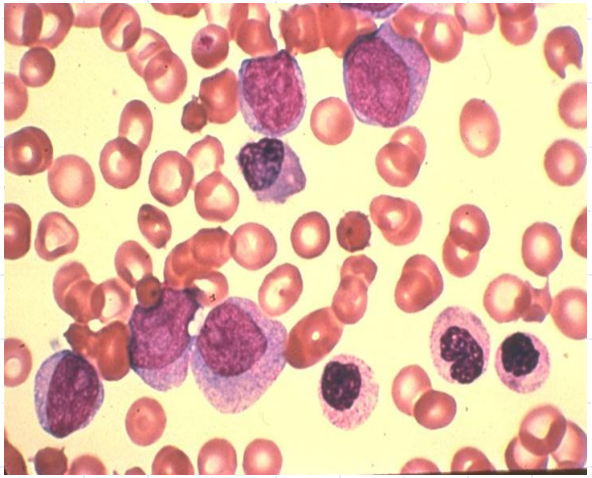
Content of Metadata

- Thesauri (Vocabularies)
  - ICD-O-03
  - ICD-9-CM
  - ICD-10
  - Online Mendelian Inheritance in Man (NCIB)
  - Unified Medical Language System (NLM)
  - SNOMED
  - …
- Key Words (Standard)
  - Acute Myelogenous Leukemia
  - Acute Myeloid Leukemia
  - Acute Non-Lymphocytic Leukemia
  - AML
  - ANLL
  - M9861/3
  - C92.0
  - 205.03
  - OMIM 6024039
  - …

World Health Organization

OMIM
Online Mendelian Inheritance in Man

SNOMED® INTERNATIONAL



Entities and attribute information

- Overview: A summary listing of the fields
  - May include field descriptions
  - Descriptive name
- Detailed Citation:
  - Descriptive and logical name
  - Data type and format
  - Allowed values or vocabulary

Example

County FIPS Code (LOC_COUNTY_FIPS)
Character, Length = 5
Values:

- 49001 = State of Utah, Beaver County
- 49003 = State of Utah, Box Elder County
- 49005 = State of Utah, Cache County
- .
- .
- .
- 49057 = State of Utah, Weber County

Logical domain

- Logical Name: HO_UBLR_STD_2007.sas
  - is the actual file name
  - this may change through data transaction services
- Operation System: LINUX
  - Allows the potential user to consider operational characteristics that may be inherited by the dataset or transaction service
- DBMS: SAS Version 9.2
  - Allows the potential user to consider DBMS characteristics that may have been imposed on the dataset or transaction service
- Location: T:EEP\EPHT\SAS\DW\UBLR\
- Record Count: 63,205 records
- File Size: 22,000 KB

Provenance

- Identity and contact information for the data owner / data steward
- Process of acquiring and compiling the data
  - Processes of transforming source data
  - Processes of deriving additional data from source data
- Processes used in preparing the data for distribution
  - Frequency
  - Quality Control

Examples

- Procedure: Data submission and registration
- Procedure: Appending electronic data
- Procedure: Standardization to PHIN-VADS
- Procedure: Manual data entry
- Procedure: Geo-coding
- Procedure: Geo-referencing
- Procedure: Aggregation
- Procedure: Data warehousing

Procedures

Include:

- Tools used (version)
- Reference data used (version)
- Description of setting, scripts, codes, etc.
- Details or mapping schema or transformation schema

Considerable posting documentation online and referencing by URL in the metadata

Completeness

- Completeness of important data elements (e.g., age, sex)
- Completeness of derivative data (e.g., geo-referencing from address data)
- If exist, measures of ascertainment

Description of data availability

- Security declaration
- Description of access requirements
- Description of use constraints
  - Use limitations
  - Required acknowledgements
  - Required over site and review
  - Data stewards rights
- Description of acquisition process
  - Agreements and certifications
  - Technological requirements for data transaction
  - Contact information
- Description of data management and final disposition requirements
- Liability disclaimer

Use constraint (example)

- **NO-USE: This data may not be used in anyway to imply data steward agency (DSA) or Utah Department of Health (UDOH) endorsement of any research objective, commercial or for-profit venture or to advertise or support a commercial product, or to direct or plan targeted advertising. This data may not be used to refute, contradict or interfere with public health policy, programs, investigations, intervention actions or health promotion activities conducted by the DSA or its agencies or any Utah State government agency or any local government public health agency in Utah. This data may not be used to identify subjects of cancer case information or the individual or organization who reported the cancer case information.**
- **PUBLICATION: The data user will comply with DSA rules for publication or presentation of this data or any results derived from this data. Publication approval of any manuscript or document must be accomplished prior to submission for publication. Data users will provide a copy of any publication draft or public presentation of this data or results derived from this data to the Utah Environmental Public Health Tracking Network (UEPHTN) which will coordinate UEPHTN and DSA approval to publish or present. See contact information in this metadata. The DSA requires 30 days to approve draft publications. The DSA will provide a response in writing to the data user.**
- **RIGHT TO REFUSAL: The DSA and/or the UEPHTN retains the right to refusal for any publication or public presentation of the data or results derived from the data.**
- **ACKNOWLEDGEMENT: Use of this data requires acknowledgement of the DSA and the UEPHTN in any publications or public presentations of the data or results derived from the data. Acknowledgement must be made that the research was supported by the DSA and by the UEPHTN, which is partially funded by the Centers for Disease Control and Prevention (CDC).**
- **AUTHORSHIP: Authorship is required when either the DSA or the UEPHTN makes substantial contribution to the data.**
- **AUDITS: The DSA and/or the UEPHTN retains the right to conduct on-site audits of the researcher with or without cause. Audits will be conducted after notification and during normal business hours by representatives of the DSA or UEPHTN. The audit will observe research practices for protecting data.**
- **REPORTS: Data users must submit annual and final reports regarding the progress and or completion of research projects to the DSA. This will be done through the UEPHTN**

Liability disclaimer (example)

- **DISCLAIMER OF LIABILITY, RELIABILITY, DAMAGES AND ENDORSEMENT. The Utah Public Health Tracking Network (UEPHTN) is maintained, managed and operated by the Environmental Epidemiology Program (EEP) within the Utah Department of Health (UDOH). In preparation of this data, every effort has been made to offer the most current, correct, complete and clearly expressed information possible.**

  Nevertheless, some errors in the data may exist. In particular, but without limiting anything here, the Utah Department of Health disclaims any responsibility for source data, compilation and typographical errors and accuracy of the information that may be contained in this data. This data does not represent the official legal version of source documents or data used to compile this data. The UDOH further reserves the right to make changes to this data at any time without notice.

  This data has been compiled by the staff of the EEP from a variety of source data, and are subject to change without notice. The UDOH makes no warranties or representations whatsoever regarding the quality, content, condition, functionality, performance, completeness, accuracy, compilation, fitness or adequacy of the data.

  By using this data, you assume all risk associated with the acquisition, use, management, and disposition of this data in your information system, including any risks to your computers, software or data being damaged by any virus, software, or any other file which might be transmitted or activated during the data exchange of this data. The UDOH shall not be liable, without limitation, for any direct, indirect, special, incidental, compensatory, or consequential damages, or third-party claims, resulting from the use or misuse of the acquired data, even if the UDOH or its agency has been advised of the possibility of such potential damages or loss.

  Format compatibility is the user's responsibility.

  Reference herein to any specific commercial products, processes, services, or standards by trade name, trademark, manufacture, URL, or otherwise, does not necessarily constitute or imply its endorsement, recommendation or favoring by the UDOH.

  The view and opinions of the metadata compiler expressed herein do not necessarily state or reflect those of the UDOH, or the data owners and shall not be used for advertising or product endorsement purposes. Use of this data with other data shall not terminate, void or otherwise contradict this statement of liability.

  The sale or resale of this data, or any portions thereof, is prohibited unless with the express written permission of the UDOH.

  If errors or otherwise inappropriate information is brought to our attention, a reasonable effort will be made to fix or remove it. Such concerns should be addressed to the EEP program manager (See Point of Contact contained in this metadata file).

Best practices

- Describes both the resources and the content
- Enhances discoverability
  - Commonly used terminology
  - Succinct complete statements
  - Rich in content
- Machine understandable
  - Hierarchical schema (objects)
  - Use of standardized search terms
  - Content syntax

Linkage ideas
(linking a metadata file to the data set)

- How to relate data file to metadata
  - By naming convention
    - HO_UBLR_STD_2007.sas
    - HO_UBLR_MD_2007.xml
  - By metadata control number
    - Datafile.1AB234CD567_89EF.sas
    - Metadata.1AB234CD567_89EF.xml
  - Becomes part of the data
- How to test for synchronization
  - Business rules
    - Does file last update date = metadata coverage end date
    - Is file record count = native dataset record count

Some opportunities

- Build on EPHTN to develop a department level standardized process for cataloging data stores
- Drives standardization and consistency across the department
  - Data architecture
  - Data availability
  - Data content

Some opportunities

- Promotes data interoperability
  - Discoverable
  - Relate-able
- Expandable
  - Documents
  - Tools
  - Organizations
  - Anything else that would benefit from cataloging

Some generalities

- Requires or promotes standardization
  - Data architecture
  - Vocabularies
- Post and reference common components
  - Standard vocabularies
  - Procedure statements
  - Access and use statements

Some questions

- Centralized versus dispersed management?
  - Consistency, continuity
  - Staying current
  - Data familiarity to describe the data
- Source data, electronic data, shareable data, all?
  - Time commitment
  - Conflicts of description
  - Different templates for different classes of data
  - Useful for data management and documenting provenance